

Influence Analysis in the Blogosphere

MICHINARI MOMMA, GREE Corp.
 YUN CHI, NEC Laboratories America
 YUANQING LIN, NEC Laboratories America
 SHENGHUO ZHU, NEC Laboratories America
 TIANBAO YANG, Michigan State University

In this paper we analyze influence in the blogosphere. Recently, influence analysis has become an increasingly important research topic, as online communities, such as social networks and e-commerce sites, playing a more and more significant role in our daily life. However, so far few studies have succeeded in extracting influence from online communities in a satisfactory way. One of the challenges that limited previous researches is that it is difficult to capture user behaviors. Consequently, the influence among users could only be inferred in an indirect and heuristic way, which is inaccurate and noise-prone. In this study, we conduct an extensive investigation in regard to influence among bloggers at a Japanese blog web site, BIGLOBE. By processing the log files of the web servers, we are able to accurately extract the activities of BIGLOBE members in terms of writing their blog posts and reading other member's posts. Based on these activities, we propose a principled framework to detect influence among the members with high confidence level. From the extracted influence, we conduct in-depth analysis on how influence varies over different topics and how influence varies over different members. We also show the potentials of leveraging the extracted influence to make personalized recommendation in BIGLOBE. To our best knowledge, this is one of the first studies that capture and analyze influence in the blogosphere in such a large scale.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications—*Data mining*

Additional Key Words and Phrases: blog, influence, link analysis, content analysis, temporal analysis

1. INTRODUCTION

Influence analysis is a very important research topic in social science and is becoming more and more important in online communities as online social networks, such as Facebook and Twitter, and online e-commerce sites, such as Amazon and Netflix, playing increasingly important roles in people's daily life. In these online communities, influence is ubiquitous: in a social network, the activities and interests of a user (e.g., what a blogger reads or writes about) are usually heavily affected by that of his or her friends in the network; in an online e-commerce site, the opinions of an authoritative reviewer can significantly sway the purchase decisions of many customers. Analyzing such influence, in addition to serving scientific research purposes, also has practical importance in various areas. For example, it may offer accurate opinion survey for politicians or play a key role in product promotion and damage control for businesses.

Social influence describes the phenomenon by which the behavior of an individual is directly or indirectly affected by the thoughts, feelings, and actions of others in a population [Kraut et al. 1998; Song et al. 2007; Cialdini 2008]. As can be seen, there are two important components in social influence. The first component is the behavior or actions of an individual, and the second component is that these actions should be a consequence of *being affected* by other people. These two components rely on a *causal* effect between the actions of an individual and that of other people that influence the individual. To detect such a causal effect in the influence is a very challenging problem. Most existing approaches adopt certain heuristics for detecting influence, e.g., by considering the temporal order of

The work of the first author was done when he was at NEC Labs.

Author's addresses: Y. Chi and Y. Lin and S. Zhu, NEC Laboratories America, Cupertino, CA 95014 USA; T. Yang, Michigan State University, East Lansing, MI, USA.

actions (user Alice is influenced by user Bob if Alice uses the same keywords [Adar et al. 2004] or the same tags [Anagnostopoulos et al. 2008] *after* Bob has done so).

These heuristics, however, failed to distinguish between the effect of *causality* and that of *correlation*. We use the blogosphere to illustrate this point. Assume blogger Alice and blogger Bob each writes a post on the topic of healthcare reform and Alice’s post dated later than that of Bob. If such actions are observed, can we draw a conclusion that blogger Alice has been influenced by blogger Bob on the topic of healthcare reform? Such a conclusion is obviously flawed because there may exist other reasons, other than Alice being influenced by Bob, for Alice and Bob to write similar posts—maybe there was a news event about healthcare reform that triggers both Alice’s and Bob’s posts. In other word, we may claim Alice and Bob are *correlated*, but should not establish *causal* relationship no matter which post is written first. Separating causality from correlation is a notoriously difficult problem [Anagnostopoulos et al. 2008]. The difficulty is partly due to that in real applications, the ground truth is usually not available in all but a few cases.¹ Without the ground truth, no one can claim they separate causality from correlation *with certain*.

Facing such a challenge, in this paper we propose a method to determine, with high level of confidence, the influence among bloggers in the blogosphere. More specifically, in this work, we investigate the influence in a close-world blogosphere, the BIGLOBE blog community Webryblog².

BIGLOBE is one of the leading Internet service providers in Japan and it provides various portal services including a blog service called Webryblog to its members. From the web server log files, we are able to capture the activities of BIGLOBE members. In this work, we mainly focus on two types of actions among BIGLOBE members: writing posts and reading other member’s posts. By studying these actions, we propose a framework to identify, with high confidence level, influence among members. Employing real actions to identify influence is a major contribution of this paper. From the identified influence, we are able to conduct in-depth analysis on how members influence each other in BIGLOBE. To our best knowledge, this is the first analysis of influence, where influence is relatively accurately identified, in such a large scale.

After obtaining the influence among its members, we are able to answer various questions about influence in Webryblog. In this paper we focus on two questions: “Are there different influential bloggers on different topics?” and “Are there different influential bloggers for different members, even on the same topic?”. Intuitively, the answers to both the questions should be *yes*. For the first question, as an anecdotal proof, if we look at the top-100 popular blog list at Techorati³, which is an authoritative blog ranking site, we can see that most of the top popular (influential) blogs only focus on a special domain (politics, technology, celebrity gossip, etc.). For the second question, we again use the previous example of healthcare reform: who are the most influential bloggers to Alice on the issue of healthcare reform probably, given Alice has Democratic leanings. To verify the above intuitions, in this paper we design several tests by leveraging some techniques we recently developed for social network analysis. From the results of these tests, we are able to provide affirmative answers to these two questions by using certain quantitative measures.

The rest of the paper is organized as follows. In Section 2, we give a survey on related work. In Section 3, we provide a detailed description of the blog data set that we use. In Section 4, we propose a method to detect influence from the user access log. In Section 5, we investigate topic-specific influence. In Section 6, we investigate member-specific influence

¹The paper citation network is such a rare case because the author of a paper usually explicitly declares the source of influence for the paper in its reference. Of course, it is not totally noise-free, due to the existence of bias [Greenberg 2009].

²www.biglobe.ne.jp and webryblog.biglobe.ne.jp.

³<http://techorati.com/blogs/top100/>.

and apply it to the application of blogger recommendation. Finally, in Section 7, we conclude and give future directions.

2. RELATED WORK

As mentioned in Section 1, the isolation of influence from other sources of correlation is known as a very challenging issue. Anagnostopoulos et al. [Anagnostopoulos et al. 2008] addressed this issue by proposing some statistical tests for isolating influence from social correlations and applied to a large data set of 340K users and 2.8M edges. Traditionally, studying similarity between people in a social network has been a central research focus. Gruhl et al. [Gruhl et al. 2004]. analyzed information propagation in the blogosphere by tracking topics in blog posts. Kumar et al. [Kumar et al. 2003] used hyperlinks to form a blog-graph and studied how the blog-graph grows over time. Adar et al. [Adar et al. 2004] introduced implicit link, or inferred link, to address the issue of sparsity of explicit URL links for the purpose of stable inference of information flow. To infer the implicit link, they used some similarity measures such as explicit URL link patterns and timing of URL link generation. They built inferred links of several thousand links from 1000 blogs. Note this approach has to rely on URL links to identify similarity between bloggers, and so it relied on *indirect* inference of implicit link.

Using instant messaging as a link medium between people, Singla and Richardson [Singla and Richardson 2008] studied social correlation and identified homophily by using demographic information as well as search queries. They showed that given a link that is defined by an instant messaging event, the probability of having the same value of user profile, including search queries or demographic attributes, is higher than the population average. In [Song et al. 2007], Song et al. proposed an information flow models where the influence is indirectly inferred by the time of adoption, e.g., innovators, early adopters, laggards, and so on.

In [Guo et al. 2009], Guo et al. studied user's posting behavior in knowledge sharing forums in detail. In online forums, Shi et al. [Shi et al. 2009] reported the probability of joining a community in terms of community features, such as size, links formed by reply-friends, and ratings of top posts. Similarity between users can explain how many common communities the users have, and it is defined by a frequency of direct reply relations and the number of common friends. In [Wang et al. 2011], Wang et al. studied and modeled how information spreads in a large enterprise throw emails. However, such information spreading is mainly task-driven, e.g., among emails about a particular consulting project. In [Romero et al. 2011], Romero et al. studied how information diffuses differently across topic among Twitter users. However, the main focus of that work is to reveal information spreading at a macroscopic level, namely over different topics, instead of at the level of influence among specific individuals. In [La Fond and Neville 2010] La Fond et al. proposed a randomization test to separate social influence (causal effects) and homophily effects (correlation). The proposed randomization test, however, is again at a macroscopic level where the aggregated edge counts related to particular attributes are used to infer influence vs. homophily.

Once influence of each user has been identified, identifying a set of most influential people would lead to interesting applications or services. The problem can be formulated as a set cover maximization problem. There has been a lot of research work for solving the problem [Leskovec et al. 2007; Richardson and Domingos 2002; Domingos and Richardson 2001; Chen et al. 2009] proposed efficient algorithms to solve the related discrete optimization problem. Arini et al. [Arini et al. 2009] addressed a personalized cover maximization problem and as an application, personal recommendation has been proposed and evaluation has been done by human subjects.

Given a graph made up from links of influential relations among bloggers, we can use link prediction techniques to predict future links. As for the link prediction *in general*, HITS [Kleinberg 1999] and PHITS [Cohn and Chang 2000] are the link analysis counter-

parts of the latent semantic analysis (LSA) that are typically used for content analysis. These methods are all based on low rank approximation of the matrix data with various interpretations from linear algebra and probability. Recently, combining link analysis with content information, for improving prediction performance, has been paid much attention. Cohn and Hofmann [Cohn and Hofmann 2000] proposed a factorization based method to incorporate both link and content information. Multi-dimensional (tensor) factorization was used by Chi et al. [Chi et al. 2009] and Chen et al. [Chen et al. 2008] for the problem. Some models that are based on the Latent Dirichlet Allocation have also been proposed [Dietz et al. 2007; Nallapati et al. 2008]. Moreover, some supervised learning methods have been developed to show promising results compared with unsupervised models [Yang et al. 2009; Lacoste-Julien et al. 2008].

3. DATA STUDY

Overall we have collected data for a period of a whole year, between September 2008 and August 2009. At Webry Blog, there are two different types of servers: one provides editing and posting service and the other browsing service. From these servers we obtained two kinds of data. One data set is a collection of blog posts, referred to as *blog content*, which is obtained from the editing servers. The other is a collection of server access logs in the same period, referred to as *access log*, which is obtained from the browsing servers.

Table I. Fields in Blog Content File and Access Log File

Fields for blog content file:

| | | | | | | | |
|------------|-----------------|--------|-----|-------|----------|------|--------|
| IP address | uploadTimeStamp | userID | URL | title | blogName | body | themes |
|------------|-----------------|--------|-----|-------|----------|------|--------|

Fields for access log file:

| | | | |
|------------|-----------------|---------|----------|
| IP address | accessTimeStamp | request | referrer |
|------------|-----------------|---------|----------|

Records obtained from the different servers are combined to conduct various analysis on influence among bloggers, which is to be described in later sections. For the purpose of matching the two data sets, we use the IP address as a key field for binding them together. Of course, use of a cookie would be a better choice because an IP address can change over time and may be shared among a number of people using the same sub-network, making the mapping between an IP address and a user many-to-many. Unfortunately, the cookie for identifying unique users was not available to us. However, we noticed in our preliminary study that by controlling the time difference, *i.e.* *window size*, of posting and browsing behaviors, the issue of shared IP address can be minimized⁴.

3.1. Blog Content

The data fields for blog posts are summarized in Table I. The raw value of IP address is hidden at BIGLOBE via a one-way hash mapping for preserving privacy.⁵ Uniqueness of the anonymized IP address remains valid so the matching operation is still valid as well. The uploadTimeStamp is the time at which a post is uploaded for the first time. Even if the blogger edits the post later, the timestamp does not change. So there is no way to tell if a post is modified after its first uploading. URL is the URL of the post, title is the title of the post, userID is identical to the domain name of the blog webpage of the blogger, blogName is the name of the series of blog posts by the blogger, body is the content of the post. Themes are the themes of the post and they are in a free description format, as

⁴We manually inspected IP addresses and blog content associated with them to estimate noise. We observed even if an IP address is shared by different bloggers, the upload time is typically different. Therefore, minimizing the window size will reduce the instances of the shared IP address.

⁵Since the reverse operation is impossible, there is no way to recover the raw value from the data.

opposed to a set of predefined categorical tags. In the editing site, the system shows popular themes from which the blogger can choose as the themes for their post, and this results in many bloggers using popular themes for their posts.

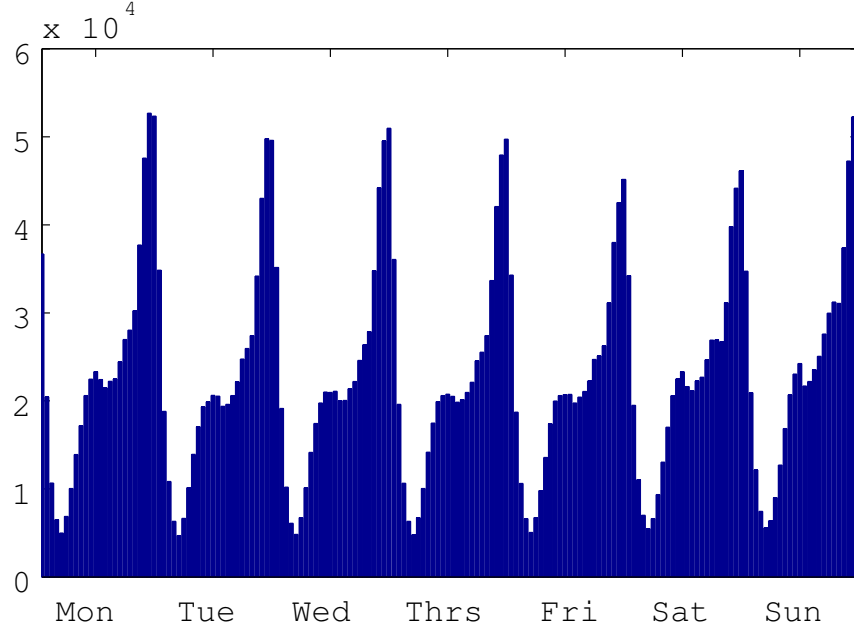


Fig. 1. The number of blog posts for week days (where actually the hourly volumes are shown).

Table II. Frequent themes: column 1 is for the rank, column 2 for ratio, and column 3 for the theme.

| Rank | Percentage(%) | Theme |
|------|---------------|---------------|
| 1 | 10.7 | diary |
| 2 | 8.00 | monologue |
| 3 | 3.41 | notes |
| 4 | 3.15 | everyday life |
| 5 | 2.21 | photograph |
| 6 | 2.10 | life |
| 7 | 1.97 | music |
| 8 | 1.66 | flower |
| 9 | 1.55 | mumble |
| 10 | 1.51 | game |
| 11 | 1.47 | gourmet |
| 12 | 1.46 | travel |
| 13 | 1.40 | movie |
| 14 | 1.32 | children |
| 15 | 1.29 | news |

During the period, the total number of blog posts was 3,870,520. The average number of post per day was 11,059. Figure 1 shows a weekly pattern of the volume of postings. As expected, a periodical trend is noticeable. Typically, Sundays have the highest volumes. Also, though not shown in the figure, holidays have higher volumes, suggesting bloggers

typically write blogs on non-working days. Qualitatively, Figure 1 has a very similar shape to the corresponding blog posting frequency distribution in [Guo et al. 2009]. This posting tendency can be explained by the popularity of diary-related blogs, since people usually write their diaries after some social events occurred and such events typically happen on Sundays or holidays. Table II shows a ranking of frequent blog themes. Note that the top four themes are all related to diaries and the total sum of them is more than 20% of the total blog posts.

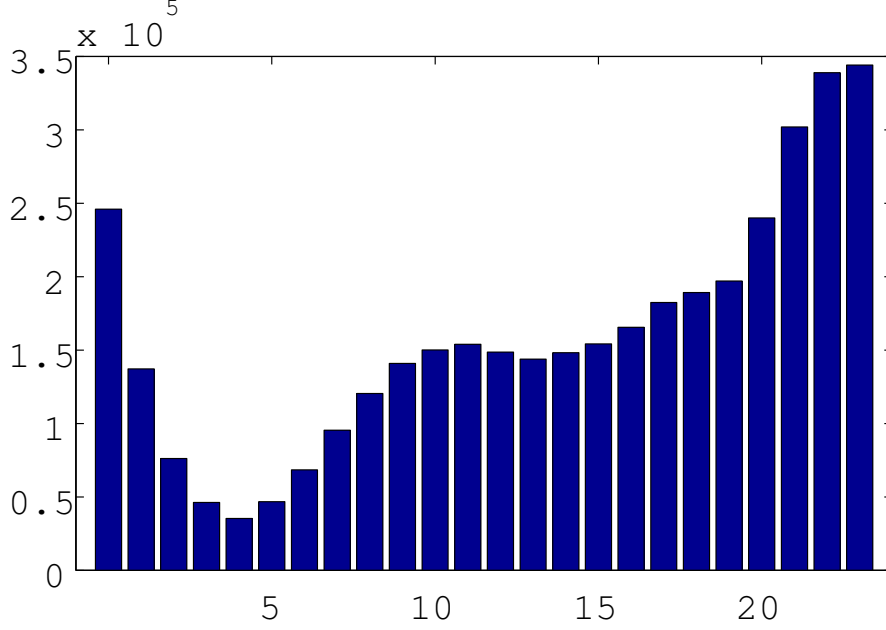


Fig. 2. The number of blog posts for each hour.

Figure 2 shows how the number of blog posts changes within a day. Not surprisingly, bloggers write blogs late at night and become less active during the daytime. Again, the shape is very similar to its counterpart in [Guo et al. 2009].

Figure 3 shows the histogram and box plot of the number of posts per blogger. The distribution is highly skewed. The mean number of posts is 37.5 while the median is 7. The skewness is due to the high volume of frequent bloggers who sit in the long tail of the distribution. Also, as revealed in the box plot, there are some bloggers who wrote extremely large number of posts.

3.2. Access Log

The format of the access log files follows the Apache combined format. Table I summarizes the fields used in our data analysis. IP address is the IP address associated with the access, accessTimeStamp is the timestamp of the access, request contains processing requests, and referrer is the referrer of the access, which is only used for removing auto-generated accesses later. In particular, to match a blog content and a record in the access log, IP address, uploadTimeStamp, and accessTimeStamp are used as a composite *soft* key.

The access logs contain all information of server accesses, which makes the number of the records massive. We remove unnecessary records to make the data processing that follows

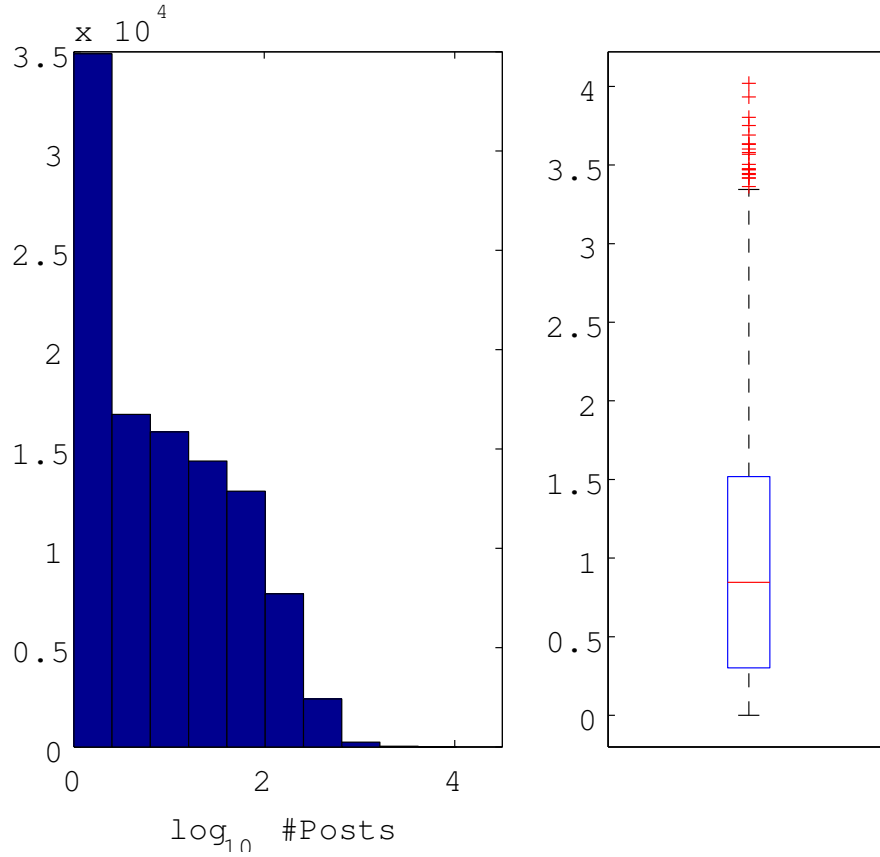


Fig. 3. *left*: Distribution of the number of posts per blogger. The number of post is on the common log scale. *right*: Box plot of the distribution.

more efficient. Since the access log is only used for analyzing bloggers' access pattern, accesses from non-bloggers are ignored. Additionally, accesses due to RSS feeds and robots are removed by using the referrer field. Further, we identify some IP addresses that are associated with anomalously large numbers of access logs. Since such accesses seem to be generated by automated measures, we remove records from such IP addresses as well. The following summarizes the criteria of removing records:

- records associated with IP address that does not appear in blog content during the data collection period
- records associated with access to their own articles
- records associated with access to index.html (contents are not available)
- records associated with access to non-html (images, etc)
- records associated with more than 12 hours before and after posting

Note the last criterion is set due to the following conflicting considerations: (1) we want to reduce noise, by minimizing the time window, because of the issue of duplicated IP address in identifying bloggers; (2) however, we do not want to restrict the time window too close to the upload time. In other words, we want to minimize the window size to remove noise but at the same time we want to maximize it to study *general* behavior of bloggers. We determine

the window size 12 hours as appropriate for the purpose of the study by inspecting and balancing the above conflicting aspects.

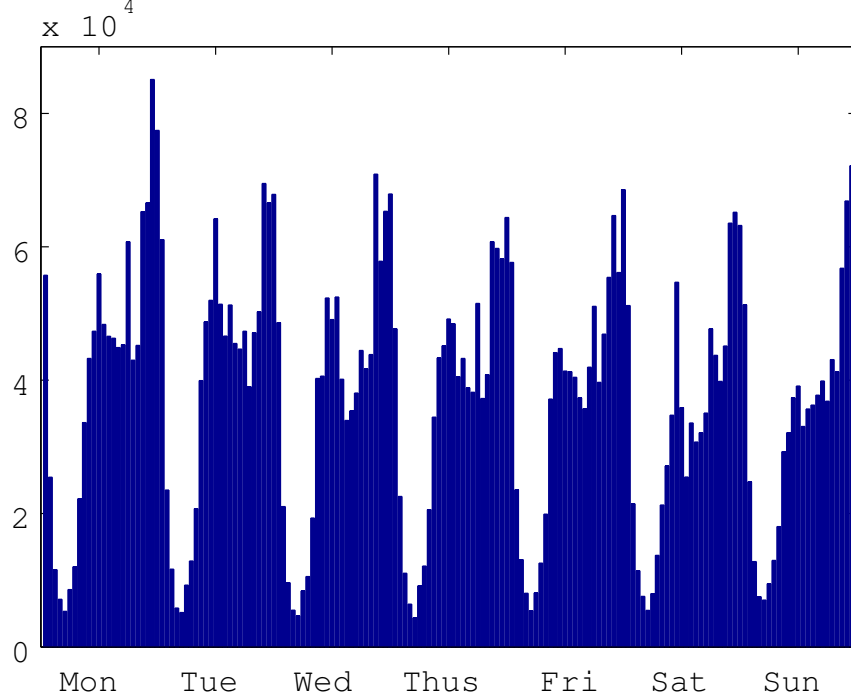


Fig. 4. The number of clicks to other blogger’s posts for each day of the week.

Figure 4 shows weekly and daily patterns of accesses. Similar to Figure 1, we can see periodic patterns of about 7 days, though the shape is not as smooth as that in Figure 2. Figure 5 shows the number of clicks on other blogger’s posts for each hour in a day. The overall trend is more uniform compared to Figure 2. Interestingly, this observation is consistent with “cut and paste” users in [Guo et al. 2009]. This suggests less serious activities on the web, such as browsing or “cut and paste” editing, follow a more uniform pattern than that for more serious activities.

4. INFLUENCE DEFINITION

As we have mentioned in Section 1, there are two components in the definition of social influence—(1) thoughts or actions and (2) the thoughts or actions should be a result of being affected by others. In this section, we investigate these two components, namely *action* and *causality*, in detail and propose a framework for identifying influence in the blogosphere with a high confidence level.

4.1. Action Selection

In our blog data set, the most frequent action is that a blogger (say A) clicks on a post (say p), which was written by another blogger (say B). However, such an action by itself is not very helpful in identifying influence among bloggers. This is because the action of clicking on p usually takes place *before* A knows about the content of p . A may have learned about p from the front-page of the web site, or she may be a loyal reader of B and therefore

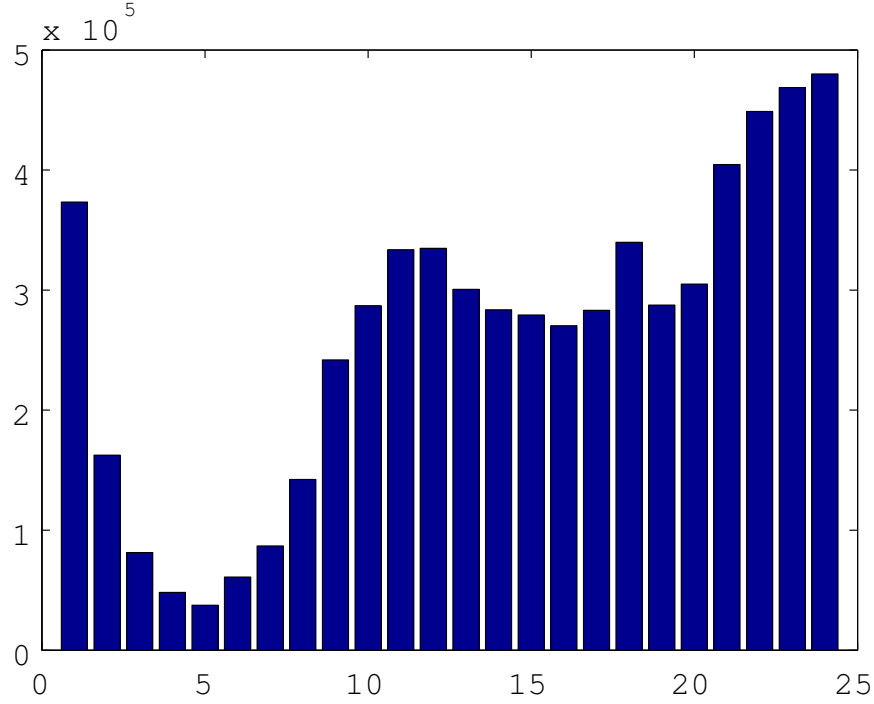


Fig. 5. The number of clicks to other blogger's posts for each hour of the day.

follows every post written by B . In either case, the action of clicking on p is not necessarily influenced by the content of p .

So instead, we focus on the action that a blogger writes a post. Because the content of a post presumably reflects its author's thoughts at the moment of writing, in the following discussion, we consider interchangeably the content of a post and the thoughts of its author at the time when the post is uploaded. While the thoughts of a blogger can be affected by many factors, such as some news she just learned about from online news sites or from TV, because we want to investigate the influence *among bloggers*, we restrict the factor to be the posts she has read before she writes her own post. For this purpose, we build a post-level network consisting of *implicit links* in the following way. We say there is an implicit link from q (written by A) to p (written by B) if A clicks on p before she writes q . That is, an implicit link (q, p) represents a high possibility that q is influenced by p . However, A may have read many posts days or months before she writes q and it is not likely all of these posts have influence on q . So to reduce noises, we adopt a time window to remove (q, p) pairs where there is too large a time gap between when A reads p and when A writes q . To select a reasonable time window, we split the time line into hourly buckets and in Figure 6(a) we plot the number of implicit links whose time gaps fall in each bucket. (To avoid the spillover to the previous day, we limit the gap to be less than 12 hours.) As can be seen, the majority of implicit links have time gaps of fewer than 5 or 6 hours. For the purpose of studying influence reflected by these implicit links, we set our time window to be 12 hours.

The access logs used in building implicit links are cleaned in the same way as that described in Section 3.2. The statistics of the resulting network of implicit link is given as:

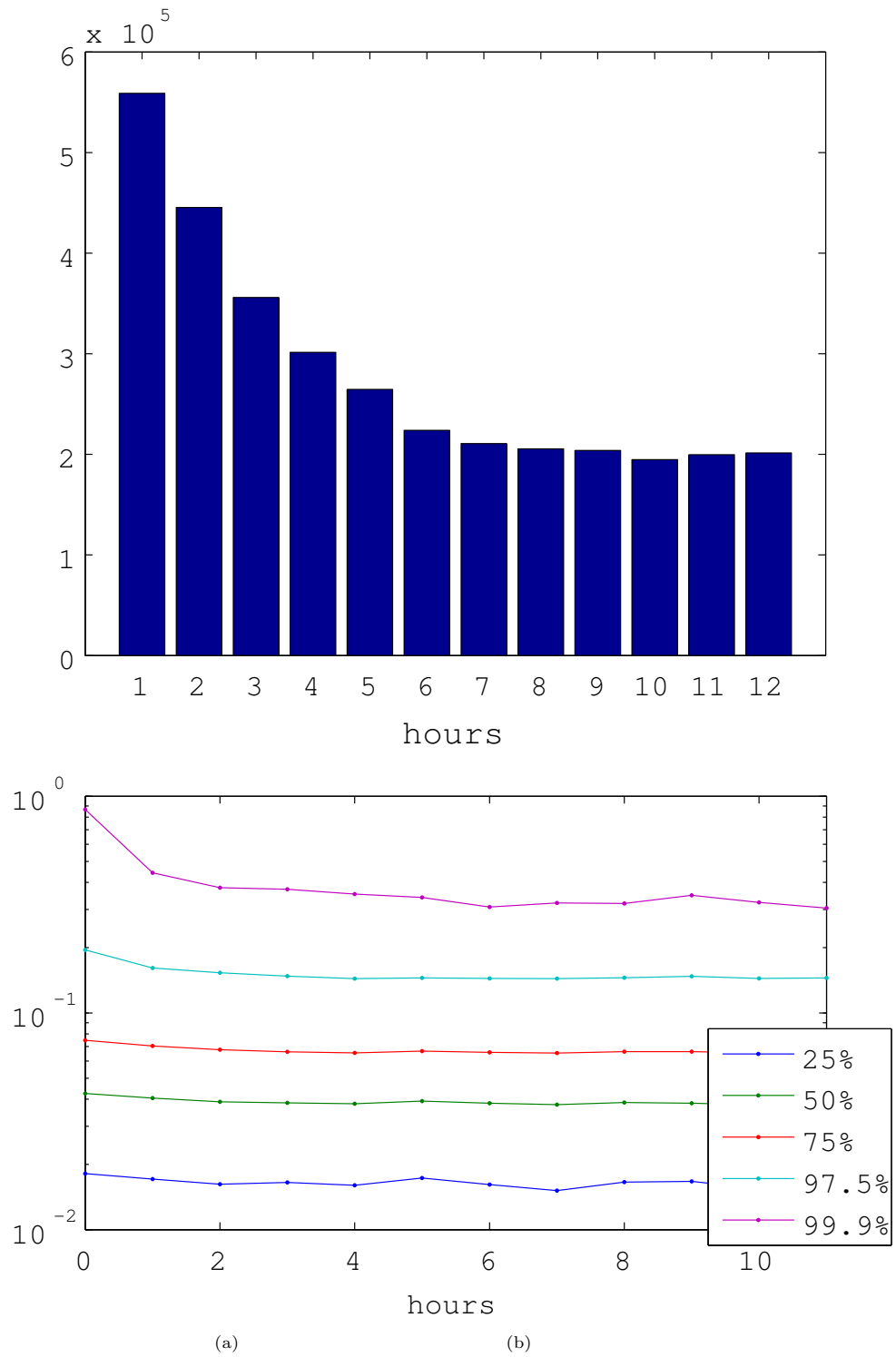


Fig. 6. (a) Volume and (b) quantile plots of implicit link 12 hours before uploading. Each point in (b) corresponding one bin in (a).

- Blogger:
 - #unique bloggers = 55,118
- Post:
 - #unique posts = 1,268,979
- Post level links (links between posts):
 - #unique implicit links = 3,847,172
- Blogger level links (links between bloggers):
 - #unique implicit links = 877,540

4.2. Causality Detection

The implicit link we just defined captures influence among bloggers to some degree—if we assume A reads each post she clicked on and each post she read influences her somewhat, both of which are reasonable assumptions, then we can infer that all the implicit links from q reflect certain influence on the content of q . However, if we simply treat implicit links as influence, we take a risk that the confidence level for the resulting influence is too low. The reason for this is again causality vs. correlation, as we will describe next.

4.2.1. Causality vs. Correlation. For each implicit link (q, p) , there can be two possible explanations for its occurrence. The first explanation is that A reads p (written by B), gets influenced, and as a result, she decides to write q . This first explanation is described by the Bayesian network in Figure 7(a), which indicates a *causal* relation between p and q . The second explanation is that A *happens to* read p and write q whereas the former takes place before the latter *by chance*. This explanation is described by the Bayesian network in Figure 7(b), which indicates a *correlated* relation between p and q . The node denoted by ? in Figure 7(b) can be some unknown reason that affects both A and B , which explains why A reads B 's posts in the first place.

As a result, in order to improve our confidence level on the influence derived from the implicit links, we need to add additional restrictions.

4.2.2. Further Criteria. We use two additional criteria to further improve our confidence level on whether an implicit link really reflects influence. We call these two criteria the *time similarity criterion* and the *content similarity criterion* and we give the definitions as follows.

time similarity. For an implicit link (q, p) to indicate influence, A must read p *shortly before* she writes q ;

content similarity. For an implicit link (q, p) to indicate influence, p must have contents *similar to* that of q .

The intuition behind the first criterion is that q is more likely to be affected by p if A reads p *shortly before* she writes q . This intuition is supported by psychology where it is well known that short-term memory and attention span for human beings are both rather short. The intuition behind the second criterion is that for p to influence q , the content of p should be somewhat *similar to* that of q . For example, A 's thoughts on healthcare reform, which she puts in q , is unlikely to be influenced by p if the content of p is about a sports game. And this is true even A reads p minutes before she writes q .

To further verify these intuitions, we conduct the following experiment to examine the content similarity between q and p for all the implicit links (q, p) 's. First, all the posts are put into the bag-of-words representation and transformed into word vectors by using a morphological analysis engine, where in the word vectors, only nouns are kept (the corpus is all in Japanese). Next, for each post q written by A , we locate all the posts $\{p_1, \dots, p_N\}$ read by A within 12 hours before q is written. Then for $\{p_1, \dots, p_N\}$, we compute their similarities $\{c_1, \dots, c_N\}$ to q , i.e., $c_i = \text{sim}(q, p_i)$. For the similarity, we choose the cosine between q and p_i . Note that to ensure the cosine values are reliable, we only compute for

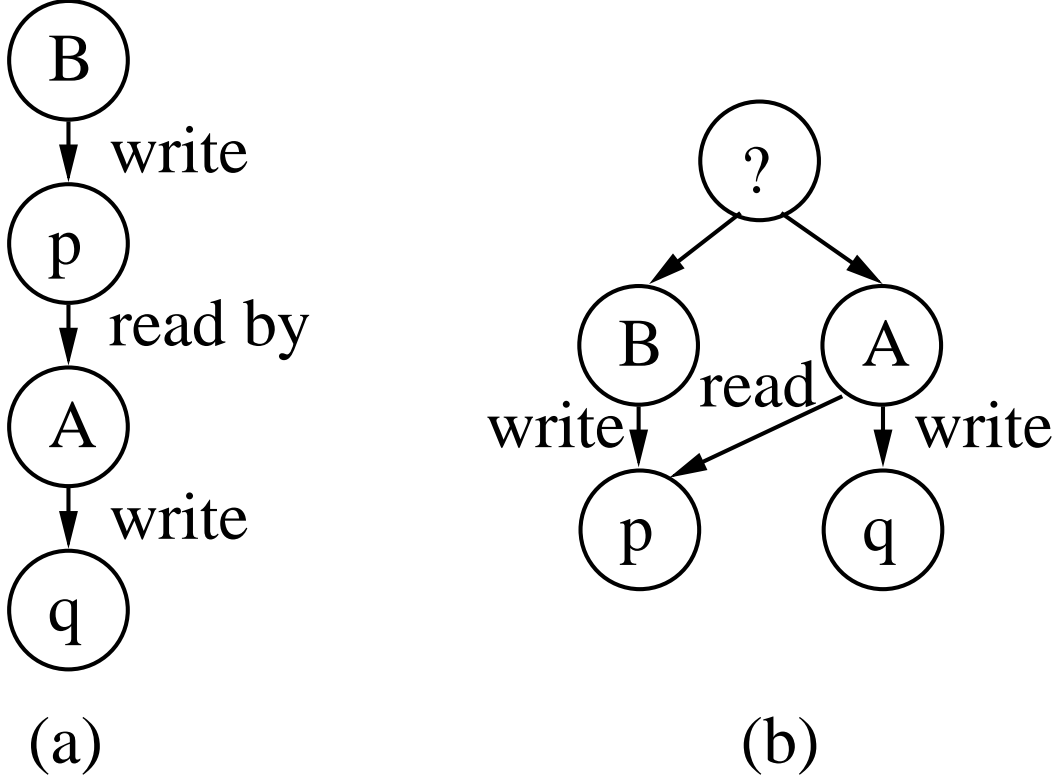


Fig. 7. Two explanations for the implicit link (q, p) , where (a) indicates a causal relationship and (b) indicates a correlated relationship.

those implicit links (q, p) where the numbers of tokens (nouns) for q and p are both at least 10. Finally, we assign to c_i the time gap (in hours) on the implicit link (q, p_i) . Figure 6(b) shows the quantile plots of cosine similarity values (of the implicit links) with time gap of 1 to 12 hours. The figure clearly shows increased similarity among implicit links with shorter time gaps. This trend is reassuring, because it shows the time similarity criterion and the content similarity criterion tend to be consistent and together they are likely to give us higher confidence level about if an implicit link actually reflects influence.

However, both the two criteria still have their problems. For the content similarity criterion, because an *absolute* similarity is used, it still suffers from the impact of correlation. To illustrate this point, we again look at the Bayesian network in Figure 7(b). Assuming the node denoted by ? is something that highly impacts the contents of p and q (e.g., it may represent some interests shared by A and B), then p and q tend to be similar whether q is influenced by p or not. Taking a more extreme example, if A reads and writes *only* about the healthcare reform and nothing else, then all the implicit links from A have high cosine similarities because of the narrow scope; in such a case, we will draw an incorrect conclusion that A is more likely to be influenced. The flaw of the time similarity criterion is that there lacks a rigorous way to tell how shortly is good enough to have causality dominating correlation. To solve these problems, we fixed the time similarity threshold and use a post-level *relative* similarity in our criteria and revise them as:

time similarity (revised). For an implicit link (q, p) to indicate influence, A must read p within τ hours before she writes q ;

content similarity (revised). For an implicit link (q, p) to indicate influence, the content of p must be more similar to that of q than an *average* implicit link from q .

We explain the two revised criteria in detail in the following.

4.2.3. Time Shuffle Test. To clarify the above two revised criteria and to make them concrete, we first describe a time shuffle test that we designed to distinguish the similarity due to influence and that due to correlation.

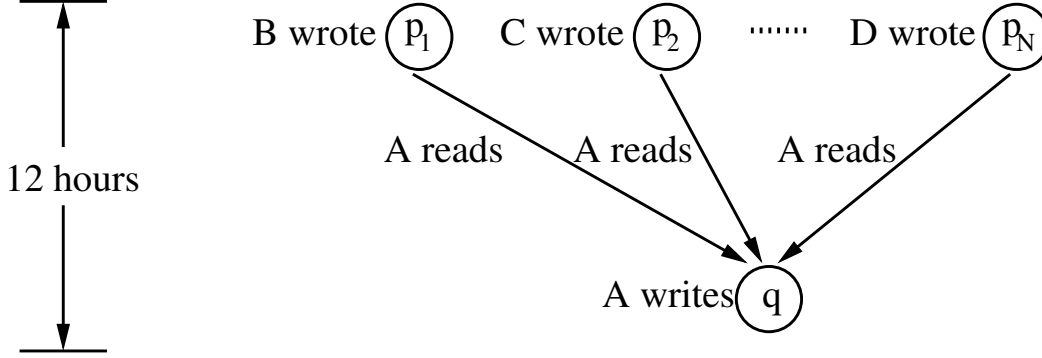


Fig. 8. Illustration of the z -test.

The basic idea of the test is based on the assumption that correlation is time-invariant (at least it is relatively homogeneous within the 12 hour time window in our data) and therefore should be insensitive to a shuffle on the time line. More specifically, for each post q , we define a fair coin in the following way. First we locate all the posts $\{p_1, \dots, p_N\}$ and compute the similarity $\{c_1, \dots, c_N\}$ in the same way as described before. After that, we find the *median* among $\{c_1, \dots, c_N\}$, which we refer to as c_k . Then we turn c_i 's into heads (H) or tails (T) according to whether c_i is greater or less than the median c_k (ties are broken randomly) to get something like $\{H_1, T_2, \dots, H_N\}$. Following that, according to the time gap on the implicit link (q, p_i) , we put H_i (or T_i) into the corresponding hour bucket. And we repeat these steps for each post, to get a series of hour buckets with certain numbers of heads and tails inside each bucket. The experimental setting of this time shuffle test is given in Figure 8.

Notice that in such a fair-coin design, instead of an *absolute* metric values, the *relative* values with respect to *the median* are used and as a result, the exact similarity metric used (cosine or Euclidean distance) is less relevant. In addition, because a fair coin is used for *each post*, the similarity bias due to the shared node, denoted by ? in Figure 7, is eliminated. To see this point, we go back to the previous extreme example: even if A only read posts about healthcare reform before she writes q , these posts are still different in terms of how similar they are to q . So we still can rank them with respect to q and obtain a fair coin. After the fair coin is obtained, the exact similarities (which may be skewed, as we discussed) are not relevant anymore.

After the buckets are collected, for each bucket we conduct a one-sample z -test. In this z -test, the statistics is the number of heads and tails in each bucket. The null hypothesis is that the bucket is generated due to correlation (therefore fair coins) and the alternative hypothesis is that the coins in the bucket are not fair. Notice that the fair-coin null hypothesis is equivalent to a fictitious test where all the p_i 's are shuffled randomly in the time line, which should give fair coins in each hour bucket. The z value for this hypothesis test is $z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$, where $\mu = 0.5$ (for the null hypothesis), \bar{X} is the sample mean (i.e.,

the fraction of heads in the bucket), σ^2 is the variance, and n is the number of samples (i.e., the number of coins in the bucket).

Figure 9(a) gives the z values for the one sample tests for the 12 hour buckets. It is obvious that the number of heads in each bucket should follow a Binomial distribution. However, because of the large number of coins in each bucket (usually tens of thousands), we can approximate the distribution accurately by a normal distribution. Under such an approximation, if we choose a p -value of 0.01, which is very typical in statistical analysis, we can reject the null hypothesis in the buckets in hours 1 and 2. It is interesting to notice that we can also reject the null hypothesis in some later hours, but with the conclusion that they are *less* similar to q than what could be explained by correlation. This effect is due to the way the coins are designed—the total number of heads is fixed (to be exactly half of all the coins) over the 12 buckets and so if the first two buckets contain more heads, the rest ones will contain less.

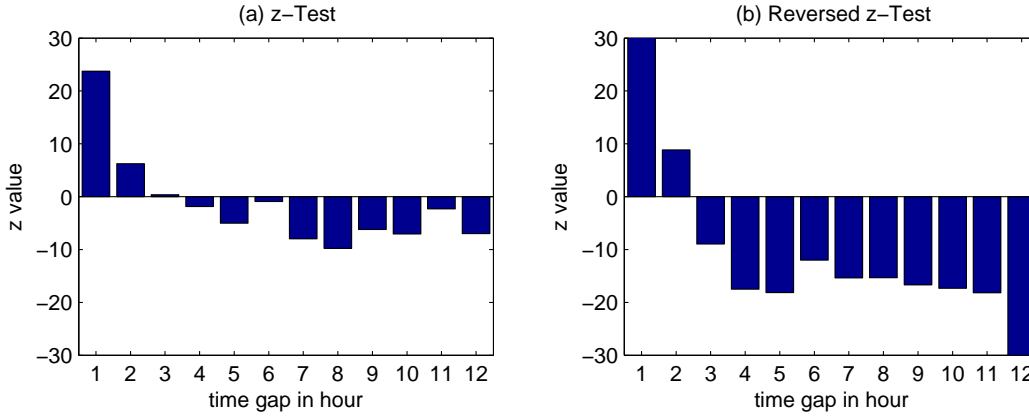


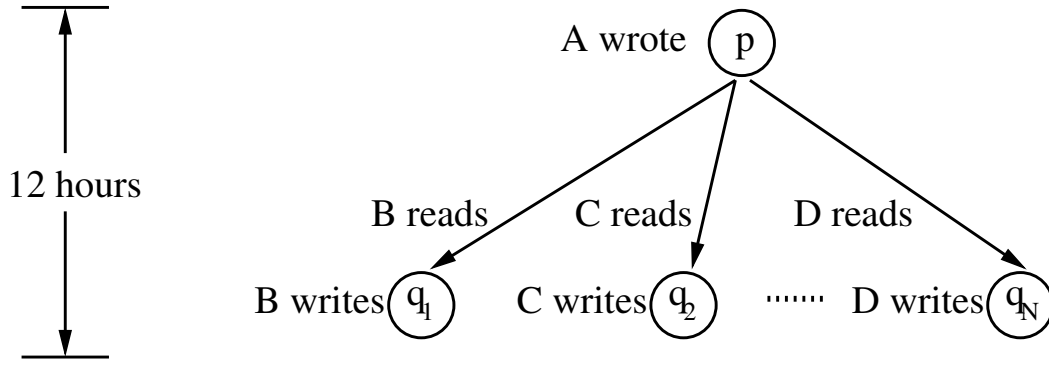
Fig. 9. Results of the hypothesis test: (a) one sample z -test, (b) reversed z -test.

4.2.4. Reversed Time Shuffle Test. In addition to the z -test just described, we also conduct a reversed z -test in the following way. We first reverse the direction of the implicit links by changing (q, p) to (p, q) and keeping the original time gap on the links. Then we conduct the same z -test on the reversed network of implicit link. In other words, this time we define a fair coin for each p instead of for each q . And we want to see among all the posts q_j 's that are written shortly after p was read, if the similarity between q_j 's and p are noticeably different for different time gaps. The experimental setting of this reversed time shuffle test is given in Figure 10.

Figure 9(b) shows the result of the reversed z -test. As can be seen, again the null hypothesis is rejected for the first two hours, except that the z values are much significant.

A possible explanation for the more significant result of the reversed z -test (in comparison to that of the z -test) is that for the z -test, the coins are defined among $\{p_1, \dots, p_N\}$, which are what A read within 12 hours, and so they are more uniform (assuming A 's interests do not change dramatically within 12 hours). In comparison, in the reversed z -test, the coins are defined among $\{q_1, \dots, q_M\}$, which are written by different bloggers over possibly much different time, and therefore they tend to be more diversified.

As a result of the time shuffle test and that of the reversed time shuffle test, we set the time τ in the time similarity criterion to be 2 hours.

Fig. 10. Illustration of the reversed z -test.

4.3. Put It All Together

With the identification of the appropriate actions and the criteria for improving confident level, we finalize our definition of influence as the following:

We say that post q (written by blogger A) is influenced by post p if (1) q is written within 2 hours after p is read by A and (2) p is more similar to q than the similarity median among all posts read by A within 12 hours before q is written.

With such a definition of influence, we are able to build, with reasonable confidence level, the influence network among posts and therefore among bloggers in our data set. The statistics of the resulting influence network is given as:

- Blogger:
 - #unique bloggers = 12,790
- Post:
 - #unique posts = 717,304
- Post level links (links between posts):
 - #unique implicit links = 487,282
- Blogger level links (links between bloggers):
 - #unique implicit links = 140,383

4.4. Further Experimental Verifications

As a sanity check of the above influence definition, we compare the top themes among all the posts and those among the posts in the influence network. Figure 11 shows the scatter plot of these top ranked themes. A theme in the upper-left corner of the figure indicates the rank of the corresponding theme is *promoted* in the influence network; a theme in the lower-right corner indicates its rank is *demoted* in the influence network. In the figure, we also show the name of several themes whose ranks dramatically changed (either promoted or demoted). As can be seen, the themes got demoted the most in the influence network are mainly about solo activities (travel, health, hobby) and those got promoted the most are mainly about group activities (baseball, PanYa—a multi-player online game, and TVXQ—about celebrity gossip). Another interesting observation we can get from the figure is that *depression seems to be rather contagious!*

Furthermore, to see how the popularity of bloggers changes in the influence network, we show the scatter plot of access rankings of influence network and implicit link network in Figure 12. Just like the theme ranking, bloggers in the upper-left corner are promoted and those in the lower-right corner are demoted in the influence network. Clearly, we can see

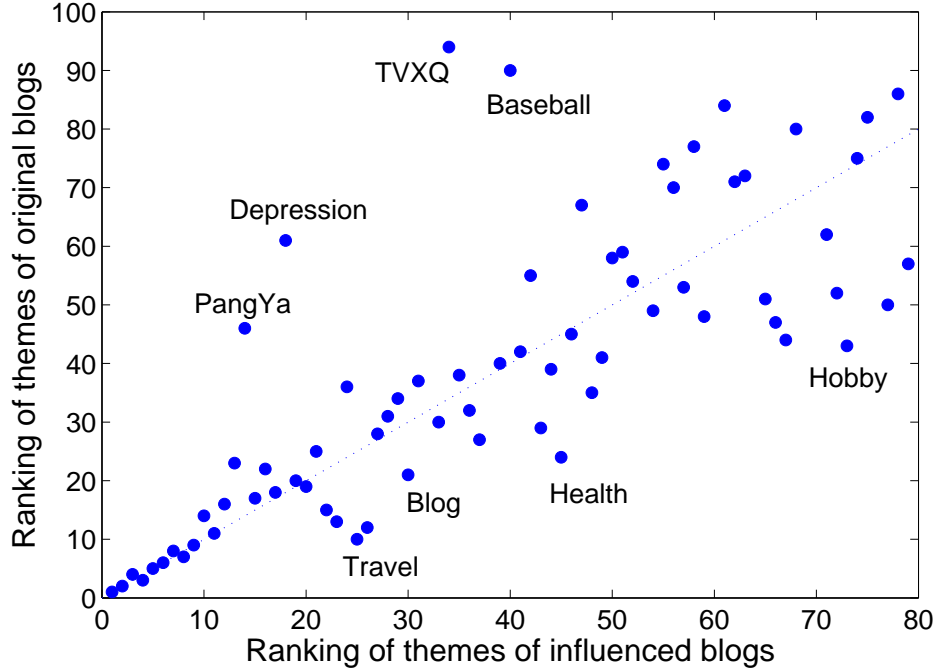


Fig. 11. Rankings of themes among original vs. influenced posts, where themes on the diagonal line have the same ranks in both the cases.

many bloggers placed above the *diagonal* line, with greater lifts in the ranks compared to Figure 11. Note the scale differences in the vertical axes. This result suggests that many bloggers who are not the most popular ones, judging by how often they are accessed, are actually very influential, judging by how often they are thought and action provoking. In the figure, we annotate some *anonymized* representative bloggers with their frequent themes (in parenthesis). As can be seen, topics such as sweets, arashi (Japanese idol group), politics, and animation are indicative of existence of *active* communities of bloggers influenced by on such themes. Note with our influence definition, such communities are formed not simply because of common interests, but because of actively participating discussions and influencing each other. On the other hand, topics such as diary, monologue are indicative of solo activities, the same as they were in the theme ranking.

5. INFLUENCE ON DIFFERENT TOPICS

After the influence is successfully extracted, in this section and the next, we analyze the influence in the Webryblog data by applying several analytical algorithms that we recently developed [Chi et al. 2009; Yang et al. 2009]. We mainly seek answers to two questions: if influence varies over different topics and if influence varies over different members. We study the first question in this section and the second question in the next section.

5.1. Influence and Topics

Are there different influential bloggers on different topics? Intuitively, the answer should be *yes*. In social science, there are several well recognized factors that determine influence,

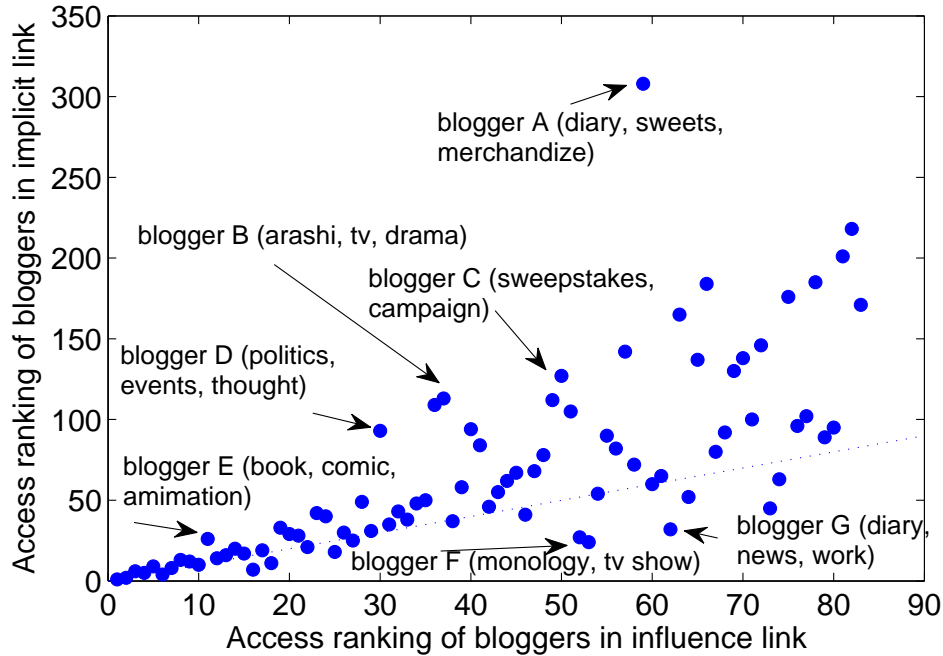


Fig. 12. Access rankings of bloggers among original vs. influenced bloggers, where bloggers on the diagonal line have the same ranks in both the cases. Blogger ID's are anonymized.

including charisma, reputation, bully pulpit, and peer pressure.⁶ Among these factors, arguably the most important one for the blogosphere is *reputation*. This is because in the blogosphere, most bloggers are ordinary people (hence charisma and bully pulpit are less important) and not necessarily know each other in the physical world (hence peer pressure is less significant). As a result, blogger *A* reads blogger *B*'s posts probably mainly due to *B*'s reputation on a given topic. For *B* to be a top influential blogger, i.e., having many readers and having these readers frequently posted responding posts, *B* must have perceived expertise and credibility. It is not likely for a blogger to have such expertise and credibility in *all* topics. This probably is the reason why for each of the top-100 most popular (influential) bloggers listed at Technorati, we can almost always assign a unique tag (such as politics, technology, celebrity gossip, etc.) to the blogger, which implies that a blogger is usually influential only on one topic.

Such an intuition sounds reasonable. However, it is a challenging problem to *quantitatively* analyze the diversity of influence of different topics. For this purpose, we analyze the influence network obtained in the previous section by using two state-of-the-art analytical algorithms. The two algorithms—iOLAP [Chi et al. 2009] and PCL-DC [Yang et al. 2009]—are recently developed by us for social network analysis. In addition, we introduce a novel metric to quantify the diversity of influence over different topics. We first give a brief overview of these two algorithms.

⁶http://en.wikipedia.org/wiki/Social_influence

5.2. Analytical Algorithms

5.2.1. iOLAP—an Approach based on Non-negative Tensor Factorization. The first algorithm, iOLAP uses non-negative tensor factorization to analyze polyadic data (those data with higher dimensions than traditional dyadic, or matrix data). For this Webryblog data set, we build the polyadic data in the following way. Assume we have determined that p (written by B) has influence on q (written by A), then for each keyword w shared by p and q (such a shared keyword always exists because the cosine similarity between p and q is nonzero according to our definition of influence), we generate a triple $\langle A, B, w \rangle$. Such a triple is a piece of evidence that A is influenced by B on w . By collecting all such triples that can be derived from the influence among all bloggers, we obtain a tensor \mathcal{D} of dimension $b \times b \times v$, where b is the number of bloggers and v is the size of the vocabulary. \mathcal{D}_{ijk} is the frequency that blogger i is influenced by blogger j on keyword k . After \mathcal{D} is constructed, we apply on \mathcal{D} the iOLAP algorithm, which seeks the optimal parameters that maximize the data log-likelihood, i.e.,

$$\arg \max_{\Theta} \sum_{i,j,k} \mathcal{D}_{ijk} \log \left(\sum_{i',j',k'} \mathcal{C}_{i'j'k'} X_{ii'} Y_{jj'} Z_{kk'} \right)$$

where the parameter Θ consists of X, Y, Z , and \mathcal{C} ; $X \in R_{b \times I}$, $Y \in R_{b \times J}$ and $Z \in R_{v \times K}$ are the major components of influencing bloggers, influenced bloggers, and key topics, respectively; and $\mathcal{C} \in R_{I \times J \times K}$ is a core tensor (of dimension $I \times J \times K$, which is much smaller than the dimension $b \times b \times v$ of the original data \mathcal{D}) that captures the interaction among the major components in X, Y , and Z . The high-level idea of the iOLAP algorithm is that, the parameter $\Theta = \{\mathcal{C}, X, Y, Z\}$ learned by the iOLAP algorithm captures I most significant groups of influential bloggers, J most significant groups of influenced bloggers, K most significant sets of topics, and the relationship among them. From the learned parameters, we are able to derive the top influential bloggers on each topic.

5.2.2. PCL-DC—an Approach based on Stochastic Block Model. The second algorithm, PCL-DC, is an improvement over the well known stochastic block model. It uses a conditional link model for link analysis and a discriminative approach to model the content. PCL-DC learns the parameters to maximize the data log-likelihood as

$$\arg \max_{\Theta} \sum_{(i \rightarrow j) \in \mathcal{E}} s_{ij} \log \sum_k y_{ik} \frac{y_{jk} b_j}{\sum_{j' \in \mathcal{LO}(i)} y_{j'k} b_{j'}}$$

where $y_{ik} = \exp(w_k^T x_i) / \sum_l \exp(w_l^T x_i)$, is the logistic discriminative model on contents. In the formula, b_i indicates the popularity (influence) of blogger i , s_{ij} is the number of links from i to j , \mathcal{E} represents the set of links, $\mathcal{LO}(i)$ represents the set of all bloggers that have influence on i , and x_i is the topic vector for blogger i . $\vec{b} \in R_b$, $Y \in R_{b \times K}$ and $W \in R_{v \times K}$ are the parameters to learn. The high-level idea of the PCL-DC algorithm is that, the parameter $\Theta = \{\vec{b}, Y, W\}$ of the PCL-DC algorithm captures following: the influence of each blogger, for each blogger how her influence is distributed among different topics, and the contents of the topics, respectively.

In addition, for the purpose of this paper, we revise the iOLAP and PCL-DC algorithms so that they share the same predefined set of topics. The main reason for this restriction is to make the results of the two algorithms comparable. The predefined K topics (the number K is, rather arbitrarily, set to 50) are obtained by using the well known PLSA algorithm. Table III shows the top keywords in several representative topics. From the keywords we can see that the topics are both unambiguous and well separated.

Table III. Top keywords (translated from Japanese) in some representative topics.

| | |
|-----|--|
| T1 | recipe, taste, salt, vegetables, ingredient, water, cake, rice, meal, meat, salad, sugar, bread, lunch |
| T2 | family, genus, leaf, flower, color, plant, seed, garden, stem, grass, name, shape, spring, autumn |
| T3 | rakuten market, shipping fee, goods, store, size, purchase, sale, price, free, item, popular |
| T4 | love, heart, hand, human, feel, meaning, life, friends, hand, world, force, image, boy friend |
| T5 | runs, starter, pitcher, game, hit, baseball, loss allowed, clutch, three strikes, batting order |
| T6 | hospital, remedy, examination, medical, doctor, pain, symptom, exercise, weight, diet, result |
| T7 | offense, gundam, enemy, harness, damage, game, recover, weapon, level, state, clear, point |
| T8 | morning, work, house, holiday, evening, dinner, lunch, shopping, yesterday, tomorrow, home |
| T9 | rail, train, station, travel, hotel, platform, express, bus, arrival, line, departure, sight seeing |
| T10 | mountain climbing, path, peaks, ridge, bifurcate, descent, departure, course, arrival, parking |

5.3. Metric for Influence Diversity

To quantitatively measure the diversity of top influential bloggers computed by iOLAP and PCL-DC on different topics, we introduce a novel metric, which we termed the *influence diversity ratio* (IDR). Here is how IDR is computed. Assume K is the number of topics. For any given positive integer N , we can compute the total number C_N of *distinct bloggers* among the top- N most influential bloggers among *all the topics*. Intuitively, a larger C_N indicates that for the fixed number of K topics, more bloggers show up in the top- N most influential bloggers and therefore the top- N influential bloggers are more diversified. C_N obviously ranges between N and KN . By normalizing C_N , we define the influence diversity ratio at N as

$$IDR_N = (C_N - N) / [N \cdot (K - 1)].$$

Notice that for any N and K , IDR_N is always between 0 and 1. $IDR_N = 0$ when the *same* set of N bloggers are ranked the top- N most influential ones among all the K topics. On the other hand, $IDR_N = 1$ when there is no overlap between the top- N most influential bloggers of *any two topics*.

5.4. Results and Discussion

Figure 13 shows the IDR_N values for $N=1$ to 50 for the most influential bloggers, derived by iOLAP and PCL-DC on the 50 topics. (Note that both the algorithms used the same set of 50 topics.) From the figure we can observe the following. First, the IDR_N values are rather high over different N and are especially high for smaller N (for N less than 10). This result verifies our intuition that the top influential bloggers among different topics should be different. Second, the IDR_N values start to decrease steadily as N grows large (for N greater than 20). This result suggests as N increases, more and more top- N influential bloggers are shared among different topics. Thirdly, comparing the IDR_N values of PCL-DC with that of iOLAP, we can see that PCL-DC gives more diversified top influential bloggers, especially for small N 's. This difference may due to the nature of the two algorithms, where iOLAP is a generative model and PCL-DC is a discriminative one.

Finally, we want to point out that we have conducted similar tests under different topic numbers K . The results, whose details are skipped due to the space limit, turn out to be similar, which implies that the exact number of topics is not a crucial factor.

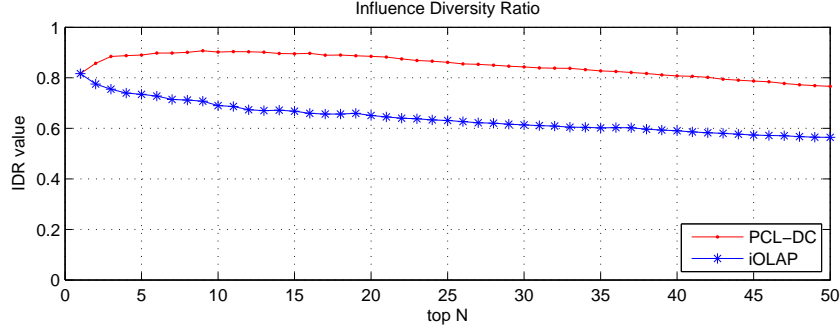


Fig. 13. The influence diversity ratio at N for iOLAP and PCL-DC results.

6. INFLUENCE ON DIFFERENT MEMBERS

In the previous section, we investigated if influence is different on different topics. In this section, we ask the question “if there are different influential bloggers for different members, even on the same topic”. Again, intuitively the answer should be *yes*. For example, on the same topic of politics, a member with Democratic leanings may be influenced by totally different people compared to a member with Republican leanings. That is, even on the same topic, different members may have different beliefs or tastes and therefore get influenced by different bloggers. To study such a *personalized* influence, of course we can use techniques similar to those used in the previous section. Instead, however, to verify that influence is personalized, we design an extrinsic test. A main reason for such a test, other than to verify influence in personalized, is to show that the extracted influence can be directly applied in practice to improve user experience in the blogosphere.

6.1. Extrinsic Test

For the extrinsic test, we use the task of personalized blogger recommendation within the given influence network. The problem is described as: *Given the historic activities of blogger A, and given a set of keywords W that A is interested in, can we recommend a blogger B, which A has not read before, that will have high influence on A on the give keywords W?* One usage scenario for this task is when a member asks the following query “*Show me top bloggers, among bloggers whose posts I have not read before, that will very likely to affect my thoughts on the issue of healthcare reform.*” For such a recommendation, we can directly use the topic-specific influential bloggers obtained in the previous section. That is, we recommend the most influential bloggers on the keywords W (we will show how to do it shortly). However, such a solution is a *global* one in that each blogger gets *the same* ranked list of influential bloggers. In contrast, by using iOLAP and PCL-DC algorithms, we are able to make *personalized* recommendation by taking into consideration the historic activities of the blogger to whom the recommendation is made.

If it turns out that the performance of the personalized recommendation is considerably better than that of global recommendation, then we can infer that personalization helps identify influential bloggers for each member, which in turn indirectly proves that influence is different for different members.

6.2. Data Preparation and Algorithms

For this recommendation task, because at the time of this writing we still are not able to plug our recommendation engine into the real-time BIGLOBE system, we simulate the recommendation task in the following way. We start by splitting the influence network into two parts—a training set and a test set. Starting from the influence network, for each node

A in the network we locate all the bloggers that have influence on A (i.e., those nodes that are reachable from node A in one step in the influence network); then among these bloggers, we randomly select one, say B , and remove link (A, B) from the influence network and put (A, B) into the test data set. After applying this process to all the nodes, the remaining influence network (with test data removed) is used as the training data set. Notice that because of the way the data are split, for each (A, B) in the test data, (A, B) is absent from the training data. By doing this, we avoid the “bookmark effect”, where A tends to read bloggers that she has read before. Furthermore, to make the influence topic-specific, we provide to the algorithms the set of keywords W on which A is actually influenced by B . That is, W is the union of overlapping keywords shared by any pair of (q, p) in the test data where q is written by A and p is written by B . In real applications, W can simply be the keywords (e.g., *healthcare reform*) provided by the blogger in her query.

For the global recommendation, we use the topic-specific influential bloggers in the following way. If B indicates a blogger, W indicates the query keywords, and t_k indicates the k -th topic, then we can write the conditional probability (keyword-specific influential blogger) of recommending blogger B given keywords W as

$$P(B|W) = \sum_k P(B|t_k) \cdot P(t_k|W)$$

where $P(B|t_k)$ is obtained in the previous section and $P(t_k|W)$ can be obtained from the output of iOLAP by using the Bayes rule: $P(t_k|W) \propto P(W|t_k) \cdot P(t_k)$. We refer this recommendation as TG, for topic-specific global recommendation.

For the personalized recommendations, for iOLAP we have

$$P(B|A, W) = P(A, B, W)/P(A, W) \propto [\mathcal{C}, X, Y, Z]_{ABW}$$

and for PCL-DC we can again use the Bayes rule.

6.3. Results and Discussion

In terms of performance metric, we use a measure typically used in the information retrieval field, *recall-at-N*. That is, if N recommendations are allowed to be made to each member, what fraction of the links in the test data are correctly recalled. Figure 14 shows the performance for all the algorithms. From the performance we can see that compared to the global recommendation (TG), the personalized recommendations (iOLAP and PCL-DC) have noticeably better performance than the global recommendation, because they considered historic behaviors and predicted the influential people for each member differently. This verifies our conjecture that influence differs over members, even on the same topic. In addition, we also show the performance of PCL, a simplified version of PCL-DC but without content analysis. The relatively poor performance of PCL demonstrates that by using link analysis alone we do not have an accurate model for the influence network. As a consequence, we have learned that a good model for the influence network relies on both contents and links. In other words, influence is both topic-specific and member-specific.

7. CONCLUSION AND FUTURE WORK

In this paper we analyzed influence in a large blog data set. We defined influence in a principled way by selecting appropriate actions, proposing intuitive criteria, and designing rigorous statistical tests. After the influence was extracted, we further investigated the questions that if influence is topic-specific and if it is member-specific. We provided affirmative answers to these questions by leveraging state-of-the-art algorithms for social network analysis, introducing novel quantitative measures, and using both intrinsic and extrinsic tests. Some of the tests also reveals the potential application of influence in blogger recommendation in the blogosphere. To our best knowledge, such an extensive analysis on influence in such a large-scale data set is the first of its kind.

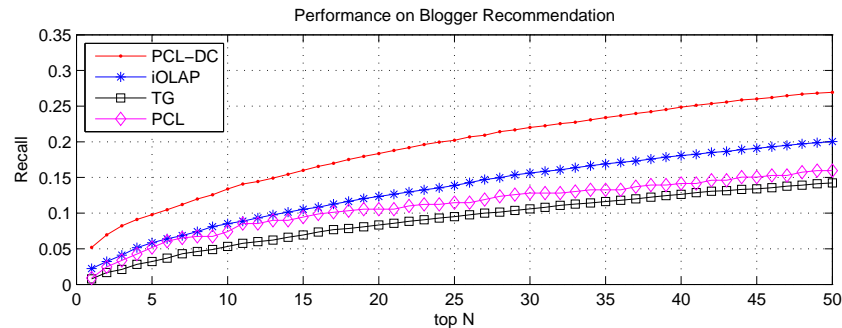


Fig. 14. Recall at top- N for the blogger recommendation task.

For future work, we plan to extend our investigation in the following two directions. First, in this paper we essentially assumed a time-invariant system where influence is considered static and topics are treated as unchanged over time. This assumption may be questionable when bloggers dynamically join and leave the system or when the topics are time-sensitive (e.g., a president election). In the future, we plan to study dynamics in this blog data set. Second, in this paper we mainly focused on one-step direct influence. But previous studies have observed cascade behavior in social networks where information is diffused through paths with multiple hops. Applying these diffusion models may lead additional insights into our analysis and is one of our future directions.

ACKNOWLEDGMENTS

The authors thank BIGLOBE for their support for preparing and providing us with the valuable data.

REFERENCES

- ADAR, E., ZHANG, L., ADAMIC, L. A., AND LUKOSE, R. M. 2004. Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem*.
- ANAGNOSTOPOULOS, A., KUMAR, R., AND MAHDIAN, M. 2008. Influence and correlation in social networks. In *KDD*. 7–15.
- ARINI, K. E., VEDA, G., SHAHAF, D., AND GUESTRIN, C. 2009. Turning down the noise in the blogosphere. In *KDD '09*. New York, NY, USA, 289–298.
- CHEN, W., WANG, Y., AND YANG, S. 2009. Efficient influence maximization in social networks. In *KDD*.
- CHEN, W.-Y., ZHANG, D., AND CHANG, E. Y. 2008. Combinational collaborative filtering for personalized community recommendation. In *KDD*. 115–123.
- CHI, Y., ZHU, S., HINO, K., GONG, Y., AND ZHANG, Y. 2009. iOLAP: A framework for analyzing the internet, social networks, and other networked data. *Multimedia, IEEE Transactions on* 11, 3, 372–382.
- CIALDINI, R. B. 2008. *Influence: Science and Practice* 5 Ed. Prentice Hall.
- COHN, D. AND CHANG, H. 2000. Learning to probabilistically identify authoritative documents. In *ICML*. 167–174.
- COHN, D. A. AND HOFMANN, T. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*.
- DIETZ, L., BICKEL, S., AND SCHEFFER, T. 2007. Unsupervised prediction of citation influences. In *ICML*. 233–240.
- DOMINGOS, P. AND RICHARDSON, M. 2001. Mining the network value of customers. In *KDD*.
- GREENBERG, S. A. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 339.
- GRUHL, D., GUHA, R., LIBEN-NOWELL, D., AND TOMKINS, A. 2004. Information diffusion through blogspace. In *WWW*. New York, NY, USA, 491–501.
- GUO, L., TAN, E., CHEN, S., ZHANG, X., AND ZHAO, Y. E. 2009. Analyzing patterns of user content generation in online social networks. In *KDD*. 369–378.

- KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5, 604–632.
- KRAUT, R. E., RICE, R. E., COOL, C., AND FISH, R. S. 1998. Varieties of social influence: the role of utility and norms in the success of a new communication medium. *Organization Science* 9, 437–453.
- KUMAR, R., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2003. On the bursty evolution of blogspace. In *WWW*. 568–576.
- LA FOND, T. AND NEVILLE, J. 2010. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*. WWW '10. ACM, New York, NY, USA, 601–610.
- LACOSTE-JULIEN, S., SHA, F., AND JORDAN, M. I. 2008. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*.
- LESKOVEC, J., KRAUSE, A., GUESTRIN, C., FALOUTSOS, C., VANBRIESEN, J., AND GLANCE, N. 2007. Cost-effective outbreak detection in networks. In *KDD*.
- NALLAPATI, R. M., AHMED, A., XING, E. P., AND COHEN, W. W. 2008. Joint latent topic models for text and citations. In *KDD*. 542–550.
- RICHARDSON, M. AND DOMINGOS, P. 2002. Mining knowledge-sharing sites for viral marketing. In *KDD*.
- ROMERO, D. M., MEEDER, B., AND KLEINBERG, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. WWW '11. ACM, New York, NY, USA, 695–704.
- SHI, X., ZHU, J., CAI, R., AND ZHANG, L. 2009. User grouping behavior in online forums. In *KDD*. 777–786.
- SINGLA, P. AND RICHARDSON, M. 2008. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW*. 655–664.
- SONG, X., CHI, Y., HINO, K., AND TSENG, B. L. 2007. Information flow modeling based on diffusion rate for prediction and ranking. In *Proceedings of the 16th international conference on World Wide Web*. WWW '07. ACM, New York, NY, USA, 191–200.
- WANG, D., WEN, Z., TONG, H., LIN, C.-Y., SONG, C., AND BARABÁSI, A.-L. 2011. Information spreading in context. In *Proceedings of the 20th international conference on World wide web*. WWW '11. ACM, New York, NY, USA, 735–744.
- YANG, T., JIN, R., CHI, Y., AND ZHU, S. 2009. Combining link and content for community detection: A discriminative approach. In *KDD*.